Math 343 / 643 Fall 2024
Midterm Examination One Solutions

Professor Adam Kapelner

February 29, 2024

Full Name _____

Code of Academic Integrity

Since the college is an academic community, its fundamental purpose is the pursuit of knowledge. Essential to the success of this educational mission is a commitment to the principles of academic integrity. Every member of the college community is responsible for upholding the highest standards of honesty at all times. Students, as members of the community, are also responsible for adhering to the principles and spirit of the following Code of Academic Integrity.

Activities that have the effect or intention of interfering with education, pursuit of knowledge, or fair evaluation of a student's performance are prohibited. Examples of such activities include but are not limited to the following definitions:

Cheating Using or attempting to use unauthorized assistance, material, or study aids in examinations or other academic work or preventing, or attempting to prevent, another from using authorized assistance, material, or study aids. Example: using an unauthorized cheat sheet in a quiz or exam, altering a graded exam and resubmitting it for a better grade, etc.

I acknowledge and agree to uphold this Code of Academic Integrity.

_____ _____

signature date

Instructions

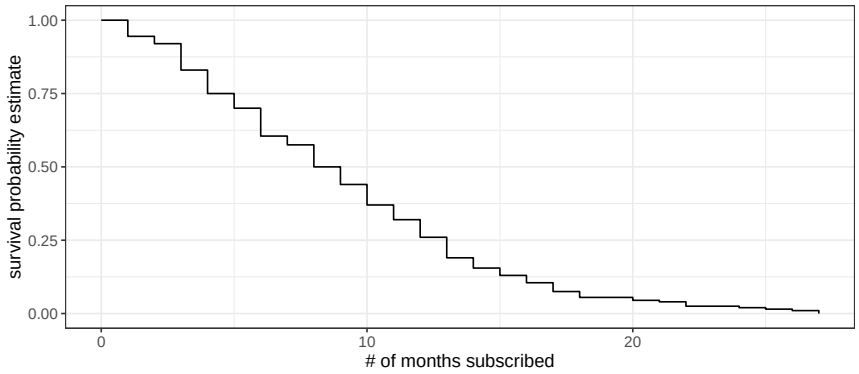
This exam is 75 minutes (variable time per question) and closed-book. You are allowed **one** page (front and back) of a "cheat sheet", blank scrap paper (provided by the proctor) and a graphing calculator (which is not your smartphone). Please read the questions carefully. Within each problem, I recommend considering the questions that are easy first and then circling back to evaluate the harder ones. No food is allowed, only drinks.

Problem 1 We are interested in understanding average churn (survival) in our monthly subscription business. Let Y model the number of complete months a customer stays subscribed. Y is discrete with $\mathbb{S}_Y = \{0, 1, 2, \ldots\}$. Zero indicates they canceled their subscription before the first month. There are $n = 200$ customers in our sample.

We take a sample today $y_1, \ldots, y_n$. Our business has been open three years, so the max survival will be $\max\{y\} = 36$. Also, people signed up at different times, so the subscriptions will be censored at all different values. Here's the raw data sorted:

```
  [1]  0   1   2   2   2   2   3   3+  3   3   4   4   5+  5   5   5   5   5   5   5   5   6
 [23]  6   6   6   6   6   7+  7   7   7   7   7   7   8   8   8   8   8+  8   8   8+  8   8+
 [45]  9   9+  9   9   9   9   9+  10  10  10  10+ 10+ 11  11  11  11+ 11  11  11+ 11  11+ 11
 [67] 11+ 11  12  12  12+ 12  12  12  12+ 12+ 12+ 12  13  13+ 13  13  13  13  13  13  13+ 13
 [89] 14+ 14  14+ 14  14+ 14  14+ 14  14  15  15+ 15  15+ 15  15  15  15  15+ 15  16  16+ 16  16
[111] 16  16+ 16  16  16  16  16  16+ 17  17  17+ 17  17+ 17+ 17  17+ 17  17  17  18  18+ 18+
[133] 18+ 18  18  18  18+ 19  19+ 19+ 19+ 19  19  19+ 19+ 20  20+ 20  20+ 20  20  20  20+ 21+
[155] 21+ 21+ 21  21+ 21  21+ 21+ 21+ 21  21+ 21  22  22+ 22+ 22  23  23+ 23+ 23+ 24  24  25+
[177] 25  25+ 25+ 25+ 26+ 26  26+ 26  27  28+ 28+ 29+ 29+ 30+ 31  32+ 32+ 32+ 32+ 32+ 35+ 36+
[199] 36+ 36+
```

The + signs indicate the specific y_i's that are censored. Let c_i denote the censoring vector which has the value 1 if censored and 0 if not censored. In total, there are $\sum_{i=1}^{n} c_i = 79$ censored observations of the total 200 observations. We first ignore censoring; i.e. we pretend $c_i = 0$ for all i. The empirical survival function $\hat{S}(y)$ is plotted below:



(a) [4 pt / 4 pts] What is the estimated probability of a customer subscribing $\geq$ 1yr?

26%

(b) [6 pt / 10 pts] Compute an approximate, asymptotically valid 95% CI for the probability of surviving >10mo to three significant digits.

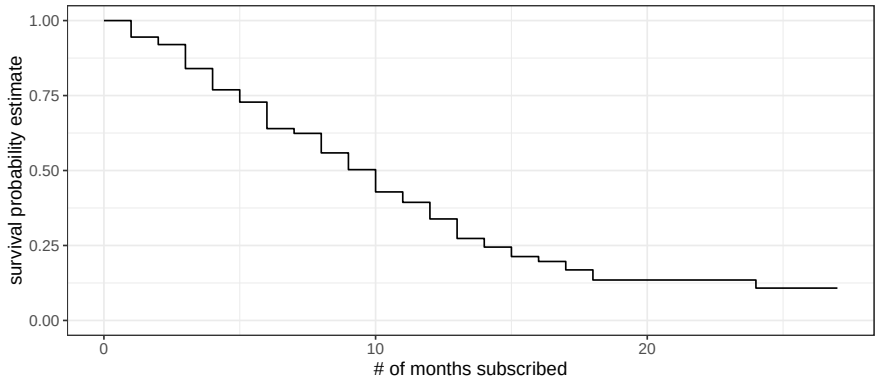
$$\left[.375 \pm 1.96 \times \sqrt{\frac{.375(1-.375)}{200}}\right] = [.341, .409]$$

(c) [4 pt / 14 pts] Is it possible to estimate the mean survival? Yes /no.

(d) [4 pt / 18 pts] Given that censoring was ignored, are the estimates in (a), (b) and (c) unbiased? Yes / No

---

We now consider censoring in estimation. The Kaplan-Meier $\hat{S}(y)$ is plotted below:



(e) [4 pt / 22 pts] What is the estimated probability of a customer subscribing $\geq$ 1yr?

35%

(f) [6 pt / 28 pts] What is the estimated median survival of a customer (in # of months)?

35%

(g) [4 pt / 32 pts] Is it possible to estimate the mean survival? Yes/ No

---

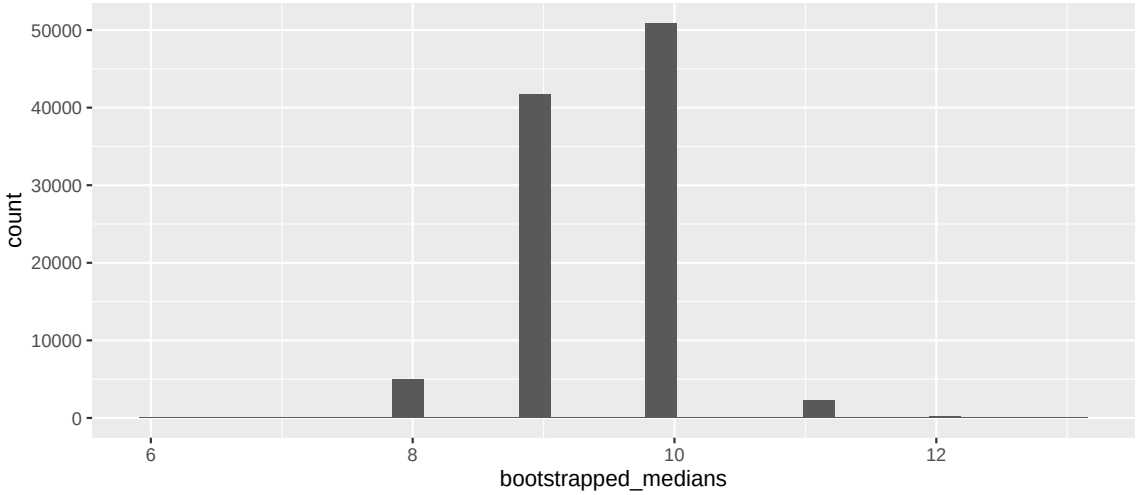
We are interested in testing H_a : MED $[Y] \neq$ 1yr. We run a bootstrap test. Below is the `R` code which makes use of the `survival` package. The objects `ys` and `c_vec` are the vectors of y_i's and c_i's respectively.

```
B = 1e5
median_bs = array(NA, B)
idx = 1 : n
for (b in 1 : B){
  idx_b = sample(idx, n, replace = TRUE)
  ys_b = ys[idx_b]
  c_vec_b = c_vec[idx_b]
  survival_fit_obj_b = survfit2(Surv(ys_b, 1 - c_vec_b) ~ 1)
  median_bs[b] = summary(survival_fit_obj_b)$table[7]
}
```

(h) [4 pt / 36 pts] How many bootstrap samples does this bootstrap test use?

100,000

Below are bootstrap samples' frequencies of the median.



(i) [6 pt / 42 pts] Consider the test $H_a : \text{MED}[Y] \neq 1\text{yr}$ at level $\alpha = 5\%$. State the test decision.

Reject H_0 / the median of Y is not equal to 1 year

We now consider the following parametric model for survival of our customers,

$$Y_1, \ldots, Y_n \overset{iid}{\sim} \text{ExtNegBinom}(\theta_1, \theta_2) := p(y) = \frac{\Gamma(y + \theta_1 - 1)}{y!\Gamma(\theta_1)}(1-\theta_2)^y \theta_2^{\theta_1},$$
$$F(y) = I_{\theta_2}(\theta_1, y+1), \quad \mathbb{E}[Y] = \frac{\theta_1(1-\theta_2)}{\theta_2}$$

where $I_a(b, c)$ is the regularized incomplete beta function. The parameter space is $\theta_1 \in (0, \infty)$, $\theta_2 \in (0, 1)$ and the support is $\mathbb{S}_Y = \{0, 1, 2, \ldots\}$.

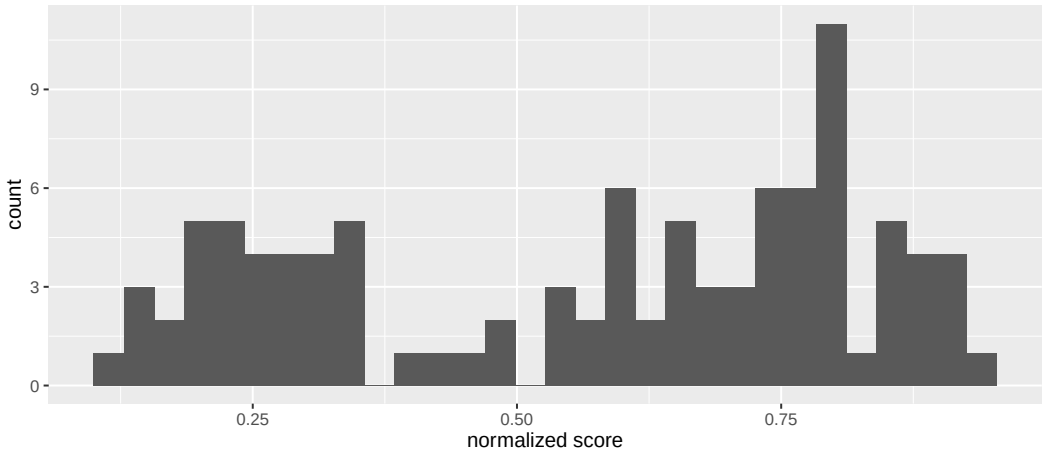
(j) [8 pt / 50 pts] Find the likelihood function. Your answer must only be a function of the data, parameter(s), the factorial function, the gamma function and the regularized incomplete beta function and fundamental constants.

$$\mathcal{L}(\theta_1, \theta_2; y_1, \ldots, y_n, c_1, \ldots, c_n) = \prod_{\{i:c_i=0\}} \frac{\Gamma(y_i + \theta_1 - 1)}{y_i!\Gamma(\theta_1)}(1-\theta_2)^{y_i}\theta_2^{\theta_1} \prod_{\{i:c_i=1\}} I_{\theta_2}(\theta_1, y_i + 1)$$

(k) [4 pt / 54 pts] Using the likelihood function above, we take its log and use an optimizer to maximize it over θ_1, θ_2. We find $\hat{\theta}_1^{\text{MLE}} = 4.128$ and $\hat{\theta}_2^{\text{MLE}} = 0.271$. Find the maximum likelihood estimate of θ, the mean subscription period measured in months to three significant digits.

$$\hat{\theta}^{\text{MLE}} = \frac{\hat{\theta}_1^{\text{MLE}}\left(1 - \hat{\theta}_2^{\text{MLE}}\right)}{\hat{\theta}_2^{\text{MLE}}} = \frac{4.128\,(1 - 0.271)}{0.271} = 11.1$$

Problem 2 We are trying to understand IQ in the army. We look at a sample of $n = 100$ IQ scores which are normalized between the minimum score (coded as 0) and the maximum score (coded as 1). No one in the sample actually has the minimum nor the maximum score. Thus we assume the scores to be $X_1, \ldots, X_n$ as iid where $\mathbb{S}_X = (0, 1)$. Here is a histogram of the data.



We believe the data has two modes so we fit the following model:

$$X_1, \ldots, X_n \overset{iid}{\sim} \theta\,\text{Beta}\,(\alpha_1,\,\beta_1) + (1 - \theta)\,\text{Beta}\,(\alpha_2,\,\beta_2)$$

and thus the likelihood and log likelihood are

$$\begin{aligned}
\mathcal{L} &= \prod_{i=1}^{n}\left(\theta\frac{1}{B(\alpha_1,\beta_1)}x_i^{\alpha_1-1}(1-x_i)^{\beta_1-1} + (1-\theta)\frac{1}{B(\alpha_2,\beta_2)}x_i^{\alpha_2-1}(1-x_i)^{\beta_2-1}\right) \\
\ell &= \sum_{i=1}^{n}\ln\left(\theta\frac{1}{B(\alpha_1,\beta_1)}x_i^{\alpha_1-1}(1-x_i)^{\beta_1-1} + (1-\theta)\frac{1}{B(\alpha_2,\beta_2)}x_i^{\alpha_2-1}(1-x_i)^{\beta_2-1}\right)
\end{aligned}$$

(a) [4 pt / 58 pts] What are the parameter(s) in this model?

$\theta, \alpha_1, \beta_1, \alpha_2, \beta_2$

(b) [4 pt / 62 pts] What type of model is this called? (The answer is one or two words).

mixture model

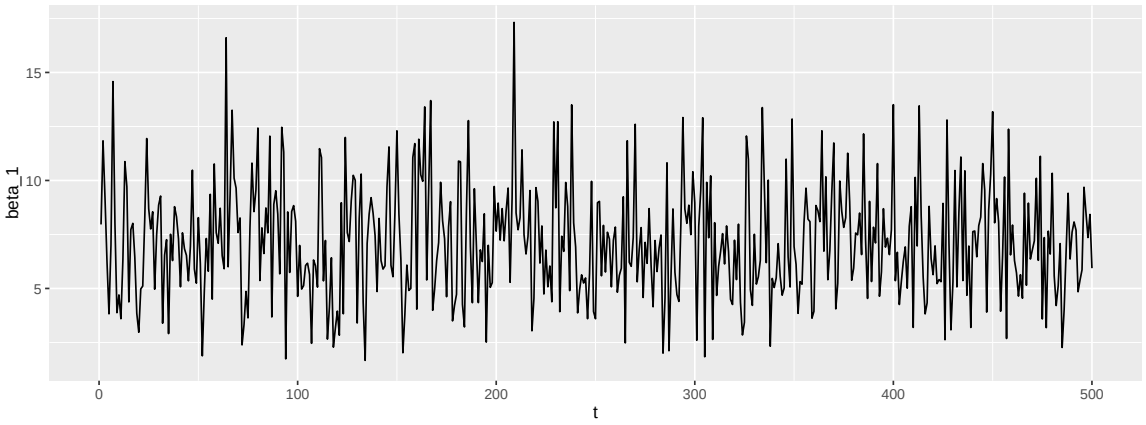
(c) [6 pt / 68 pts] If you were to use data augmentation, how would you define I_i?

$$I_i = \mathbb{1}_{\text{the observation } i \text{ belongs to the first distribution, Beta}(\alpha_1, \beta_1)}$$

(d) [4 pt / 72 pts] After data augmentation and assuming flat priors on all parameters, you use MCMC to estimate the parameters. The α_1 sampling step would be a ...

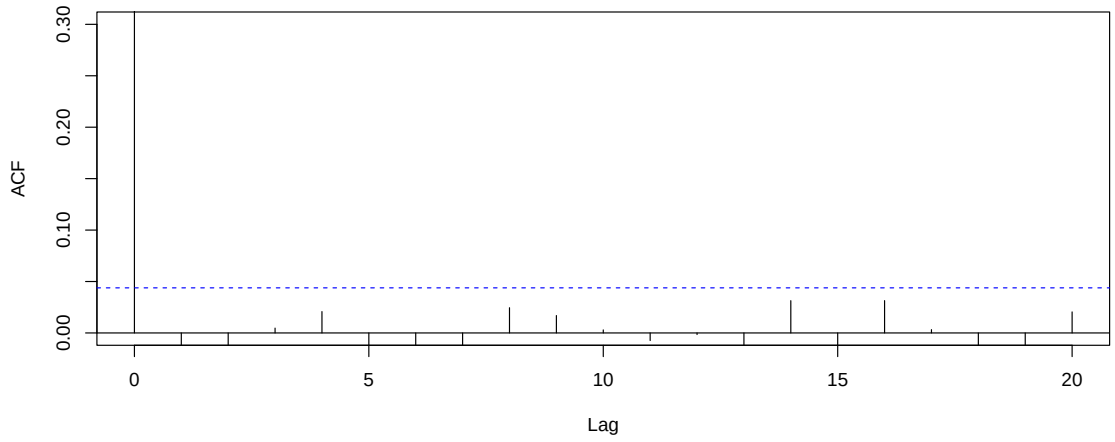
circle one of the two choices: Gibbs step / Metropolis-Hastings step.

We now implement an MCMC for inference using the incredible `stan` software. Below is the first 500 samples in the β_1 chain:



(e) [4 pt / 76 pts] Would this chain need to be burned in? Yes / No

Here is the autocorrelation plot for the β_1 chain.

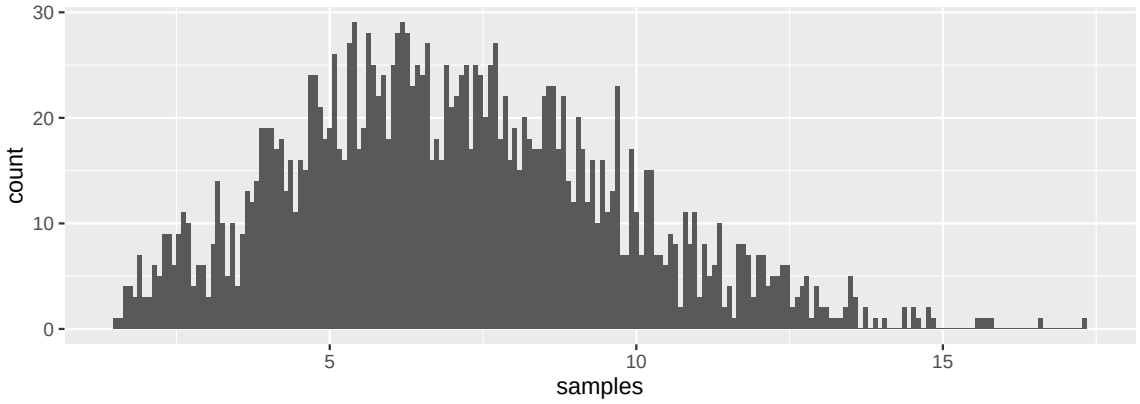


(f) [4 pt / 80 pts] At what number of samples should this chain be thinned? If the chain does not need to be thinned, write "1" below.

1

(g) [4 pt / 84 pts] After burning and thinning, are the samples considered iid? Yes /no

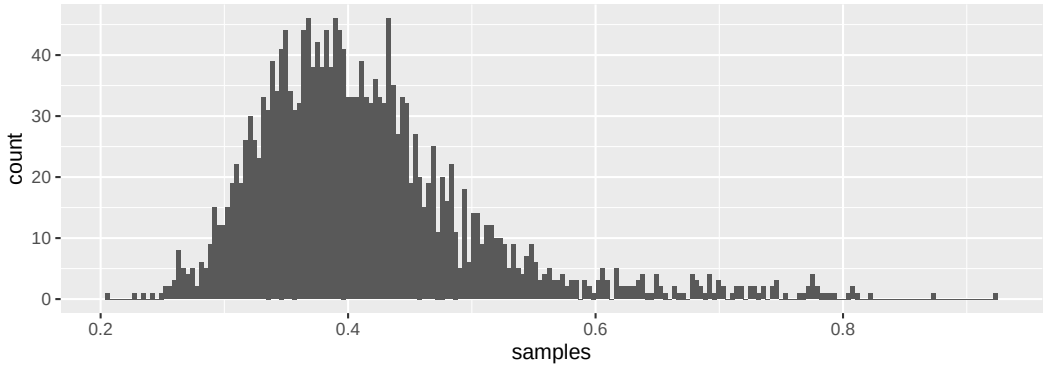
Here is the histogram of the β_1 chain after burning and thinning:



(h) [4 pt / 88 pts] Estimate the MMSE for β_1.

7

Here is the histogram of the θ chain after burning and thinning:



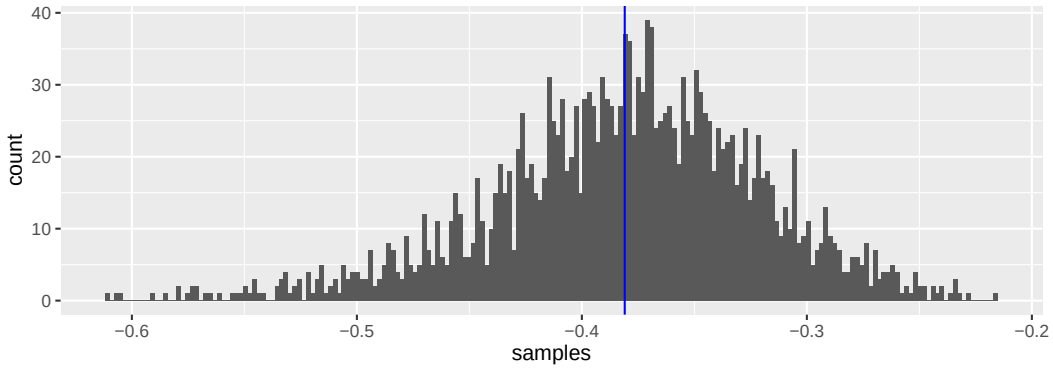
(i) [6 pt / 94 pts] Estimate $CR_{\theta,95\%}$.

$[0.275, 0.725]$

We are now interested in the difference of the two modes we observe in the data. Recall the mode of a beta distribution is given by $(\alpha - 1)/(\alpha + \beta + 2)$ if $\alpha, \beta > 1$. We are very confident that both $\alpha_1, \beta_1, \alpha_2, \beta_2 > 1$. We create a new parameter,

$$\tau := \frac{\alpha_1 - 1}{\alpha_1 + \beta_1 + 2} - \frac{\alpha_2 - 1}{\alpha_2 + \beta_2 + 2}.$$

Using the MCMC samples, we compute the τ's and show a histogram of this new *derived* chain after burning and thinning. The vertical line indicates the $\hat{\tau}^{MMSE}$:



(j) [6 pt / 100 pts] Consider the test "H_a : the two modes are unequal" at level $\alpha = 5\%$. State the test decision

Reject H_0 / the two modes are unequal